CMCS Highlight: Data Pedigree

http://cmcs.ca.sandia.gov/

January, 2003

Pedigree Browsing Concept Demonstrated in CMCS Portal

The CMCS project has demonstrated a portal-based mechanism for searching and browsing the pedigree of chemical science data as part of a general metadata and pedigree management capability. Data pedigree, sometimes referred to as data provenance, documents the inputs required to create a data set, thus providing a "line of ancestors". This allows for the categorization and tracing of the origins of scientific data, within projects and potentially across chemical scales a to the data's ultimate origins in experimental measurements or theoretical calculations. CMCS has defined a core set of pedigree relationships such as "has inputs" and "is part of" and has adopted the Dublin Core schema for pedigree related metadata such as "creator", "creation date", "publication date". CMCS has also defined a core set of chemistry-related metadata such as "chemical formula", "Chemical Abstracts Service (CAS) number", and various chemical properties. Additional pedigree relationships and metadata can be defined by users as desired.

The Pedigree Browser

All pedigree relationships and metadata are stored in CMCS's webDAV-based data repository as properties. As shown in Figure 1, the CMCS Pedigree Browser allows configurable subsets of this information to be visualized within the CMCS portal. The

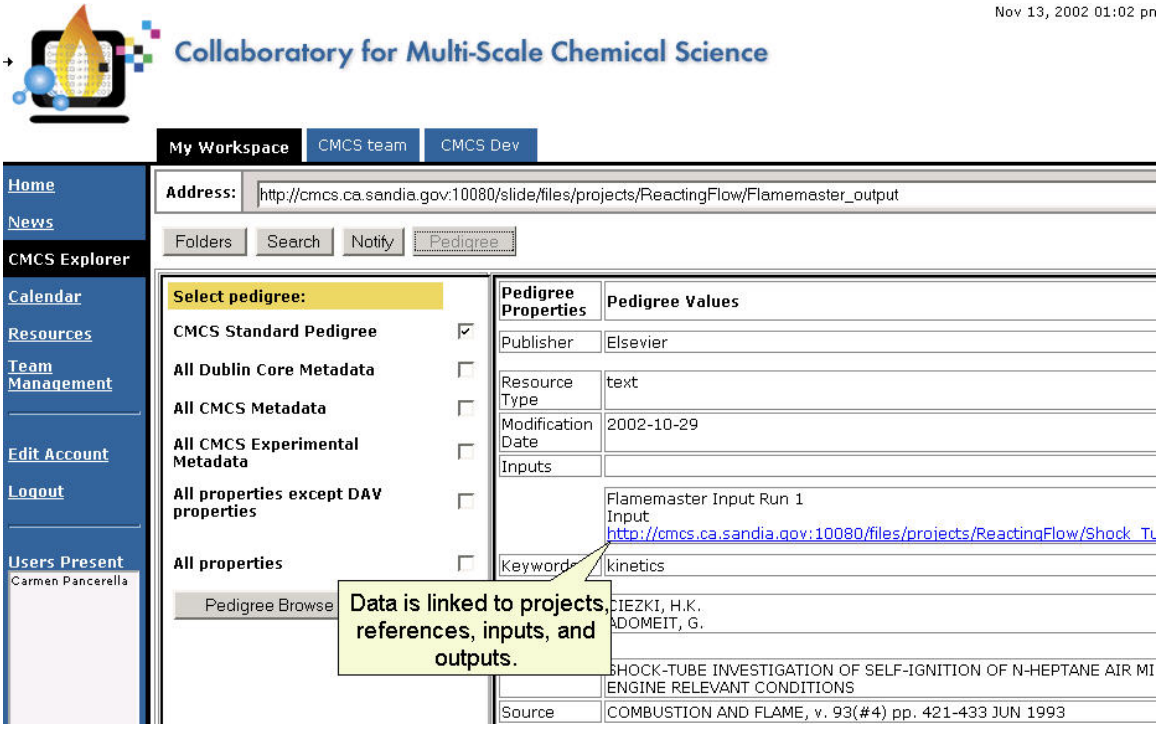

Figure 1. A screenshot of the Pedigree Browser running in the CMCS Portal. The pedigree properties shown include a link to the Flamemaster input files used to create the data, and information about the paper in which the data were published.

Browser is integrated with other portal tools and can be used to show the pedigree of items in shared project folders or search results. The Browser displays relationships as live HTML links, enabling users to quickly follow pedigree relationships, browsing, for example, from a data object to one of its inputs and from there to information about the program used to create that input.

Pedigree Examples

Hundreds of data sets of many different types, representing several chemistry-related databases, have been annotated with pedigree properties and included in demonstrations at the SC 2002 Conference and elsewhere. For example, browsing a data file returned by a search query, one might discover that it was developed as part of the Gas Research Institute Mechanism (GRI Mech) Project GRI Mech Project. One could then browse to the project web page, to the overall GRI Mech data collection and other project data, and to the literature references for the data. The demonstration with the longest pedigree data trail involves following GRI Mech pedigree links to underlying Active Tables thermochemical data from which they derive, and links from these data to the Ecce/NWChem computations of molecular properties used to create them. This demonstrates the ability to track information to its ultimate origin, across physical scales and scientific applications. Another pedigree demonstration shows how electronic notebook entries about the data will also appear within the Browser as an additional type of pedigree information.

Automating the Generation of Pedigree Data

To be visible to CMCS tools, pedigree data must be stored as webDAV properties. Properties can be generated manually using any webDAV browser, e.g. DAV Explorer, or, in the near future, through a CMCS web form. Alternately, applications can record pedigree properties directly using the webDAV protocol and/or by using the CMCS pedigree Java API, as in done in the Extensible Computational Chemistry Environment (ECCE) and the Electronic Laboratory Notebook (ELN). The final, and most transparent, option is to define how properties should be created from file content during upload to the CMCS repository. This capability is derived from CMCS's use of the Scientific Annotation Middleware (SAM) webDAV service. SAM allows registration of XML-based binary format description (BFD) and XSLT templates that describe how information should be extracted from binary, ASCII, and XML files to automatically generate properties.

External collaborations and interactions

In developing these capabilities, the CMCS team has collaborated closely with the SAM project and has held ongoing discussions with Earth Systems Grid II team members. In addition, Carmen Pancerella, Larry Rahn, and Jim Myers presented CMCS pedigree concepts and participated in general metadata discussions at the Workshop on Data Provenance and Data Derivation, Chicago, IL, October, 2002.

Point of contact for this highlight:

Carmen Pancerella
Sandia National Laboratories
carmen@ca.sandia.gov
617-630-0316